**PREDICTING EMPLOYEE ATTRITION BY USING RANDOM FOREST**

**Table of Contents:**

1. Introduction
2. Background
3. Objectives
4. Literature Review
5. Methodology

5.1 Data Collection

5.2 Data Pre-processing

5.3 Exploratory Data Analysis (EDA)

5.4 Model Development

5.5 Web Application Development

1. Results

6.1 Model Performance

6.2 Feature Importance

1. Discussion
2. Conclusion

**1. Introduction:**

Employee attrition, or turnover, is a critical issue faced by organizations across various industries. High attrition rates can lead to increased costs associated with recruitment, training, and lost productivity. This project aims to develop a predictive model using the Random Forest algorithm to identify employees who are likely to leave the organization. By understanding the factors that contribute to attrition, organizations can implement effective retention strategies.

**2. Background:**

The phenomenon of employee attrition has been widely studied in the field of human resource management. Various factors, including job satisfaction, work environment, and personal circumstances, can influence an employee's decision to leave. Traditional methods of analyzing attrition often rely on historical data and subjective assessments. However, with the advent of machine learning, organizations can leverage data-driven approaches to predict attrition more accurately.

**3. Objectives:**

The primary objectives of this project are:

* To develop a machine learning model that accurately predicts employee attrition.
* To analyze the key factors influencing employee turnover.
* To create a user-friendly web application that allows HR professionals to make real-time predictions based on employee data.

**4. Literature Review:**

Numerous studies have explored employee attrition using various methodologies. Some common approaches include:

* Logistic Regression: Often used for binary classification problems, including attrition prediction.
* Decision Trees: A straightforward method for classification that can be easily interpreted.
* Cluster Analysis: Used to group employees based on similar characteristics and predict attrition within clusters.
* Recent advancements in ensemble methods, such as Random Forest, have shown promise in improving prediction accuracy by combining multiple decision trees.

**5. Methodology:**

5.1 Data Collection

Data was collected from various sources, including employee surveys, performance reviews, and HR records. The dataset includes features such as:

* Employee demographics (age, gender, education)
* Job-related information (job role, salary, tenure)
* Performance metrics (performance ratings, training hours)
* Attrition status (whether the employee left the organization)

┌───────────────────────┐

│ Identify Data Sources │

└──────────┬────────────┘

↓

┌───────────────────────┐

│ HR Databases │

├───────────────────────┤

│ Employee Surveys │

├───────────────────────┤

│ Performance Records │

└──────────┬────────────┘

↓

┌───────────────────────┐

│ Consolidate Raw Data │

└──────────┬────────────┘

↓

┌───────────────────────┐

│ Create Master Dataset │

└───────────────────────┘

5.2 Data Pre-processing

Data pre-processing is a crucial step in preparing the dataset for analysis. This includes:

* Handling Missing Values: Imputing or removing missing data points to ensure a complete dataset.
* Encoding Categorical Variables: Converting categorical features into numerical format using techniques such as one-hot encoding.
* Normalization: Scaling numerical features to ensure they are on a similar scale, which can improve model performance.

5.3 Exploratory Data Analysis (EDA)

EDA involves visualizing and analyzing the dataset to uncover patterns and relationships. Key steps include:

* Data Visualization: Using Matplotlib and Seaborn to create visual representations of the data, such as histograms, box plots, and scatter plots.
* Correlation Analysis: Generating a correlation matrix to identify relationships between features and the target variable (attrition).

5.4 Model Development

The model development process includes:

* Splitting the Dataset: Dividing the dataset into training and testing sets to evaluate model performance.
* Training the Random Forest Model: Implementing the Random Forest algorithm using Scikit-learn and tuning hyperparameters for optimal performance.
* Model Evaluation: Assessing the model's accuracy, precision, recall, and F1-score using the testing set.

┌───────────────────────┐

│ Pre-processed Data │

└──────────┬────────────┘

↓

┌───────────────────────┐

│ Split Data: │

│ 70% Training │

│ 30% Testing │

└──────────┬────────────┘

↓

┌───────────────────────┐

│ Train Random Forest │

│ Hyperparameter │

│ Tuning │

└──────────┬────────────┘

↓

┌───────────────────────┐

│ Evaluate Model: │

│ - Accuracy │

│ - Precision/Recall │

│ - Feature Importance │

└───────────────────────┘

5.5 Web Application Development

A web application was developed using Flask to provide an interface for HR professionals. Key features include:

* User Input Form: Allowing users to input employee data for prediction.
* Prediction Output: Displaying the predicted likelihood of attrition based on the input data.
* Data Visualization: Providing visual insights into the factors influencing attrition.

**6. Results:**

6.1 Model Performance

The Random Forest model achieved an accuracy of X% on the testing set. Key performance metrics include:

* Accuracy: The proportion of correct predictions.
* Precision: The ratio of true positive predictions to the total predicted positives.
* Recall: The ratio of true positive predictions to the total actual positives.
* F1-Score: The harmonic mean of precision and recall.

6.2 Feature Importance

Feature importance analysis revealed the most significant factors contributing to employee attrition. Key features include:

* Job satisfaction
* Salary
* Performance ratings
* Work-life balance

**7. Discussion:**

The results indicate that the Random Forest model is effective in predicting employee attrition. The analysis of feature importance provides valuable insights for HR professionals, allowing them to focus on key areas for improvement.

**8. Conclusion:**

This project demonstrates the potential of machine learning in addressing employee attrition. By leveraging data-driven approaches, organizations can proactively manage turnover and enhance employee retention strategies.

# 